Journal of Digital Imaging (2018) 31:768–775
https://doi.org/10.1007/s10278-018-0096-5

Modeling Human Perception of Image Quality

Oleg S. Pianykh [1,2] · Ksenia Pospelova [2] · Nick H. Kamboj [3,4]

Published online: 2 July 2018

Abstract
Humans can determine image quality instantly and intuitively, but the mechanism of human perception of image quality is unknown. The purpose of this work was to identify the most important quantitative metrics responsible for the human perception of digital image quality. Digital images from two different datasets—CT tomography (MedSet) and scenic photographs of trees (TreeSet)—were presented in random pairs to unbiased human viewers. The observers were then asked to select the best-quality image from each image pair. The resulting human-perceived image quality (HPIQ) ranks were obtained from these pairwise comparisons with two different ranking approaches. Using various digital image quality metrics reported in the literature, we built two models to predict the observed HPIQ rankings, and to identify the most important HPIQ predictors. Evaluating the quality of our HPIQ models as the fraction of falsely predicted pairwise comparisons (inverted image pairs), we obtained 70–71% of correct HPIQ predictions for the first, and 73–76%for the second approach. Taking into account that 10–14% of inverted pairs were already present in the original rankings, limitations of the models, and only a few principal HPIQ predictors used, we find this result very satisfactory. We obtained a small set of most significant quantitative image metrics associated with the human perception of image quality. This can be used for automatic image quality ranking, machine learning, and quality-improvement algorithms.

Keywords Image quality assessment · Elo rating · Linear regression · Entropy · Fractal dimension · Gaussian pyramid

Introduction

Humans can glance at an image and instantly conclude its quality as "bad," "perfect," "too noisy," or "too dark." However, can we quantify this intuitive perception, and can we build numerical models to approximate its value?

The problem of objective image quality assessment can be found in many applications, ranging from simple web browsing to the most sophisticated machine and pattern learning. We have faced this problem many times through our work in medical imaging. In radiology, the quality of medical images has direct impact on the quality of diagnosis and patient treatment. Moreover, our ability to increase image acquisition quality in medicine often bounded by the harmful side effects (such as radiation in computed tomography, CT). To overcome this limitation, a number of image quality enhancing algorithms have been suggested and implemented—CT low-dose image enhancement being one of them [1]. However, the choice and the magnitude of this enhancement have been always left to humans, very subjective decision makers as they are. If an objective HPIQ "calibration" were possible, subjectivity could be completely eliminated from the radiology QA process.

Finally, the idea of objective image quality calibration is not foreign to medicine, and has been already tried in some basic ways: calibrating imaging scanners for SNR, or calibrating medical displays for DICOM greyscale. As a result, the goal of this study was to determine the most important numerical metrics that can be used to model HPIQ.

To achieve this goal, we considered HPIQ as a multivariate characteristic perceived by an average human observer looking at a given image. We also viewed HPIQ as a general, task-independent property the human visual system—similar to recognizing color or contrast. Therefore, we hypothesized

✉ Oleg S. Pianykh
opiany@gmail.com

Ksenia Pospelova
nephidei@gmail.com

Nick H. Kamboj
nick.h.kamboj@astonjamesllc.com

[1] Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

[2] National Research University Higher School of Economics, Moscow, Russia

[3] Aston & James, LLC, Chicago, IL, USA

[4] Harvard Extension School, Boston, MA, USA

that despite inevitable observer variability, HPIQ can be approximated as an objective (observer-independent), task-independent, and absolute (non-reference independent on the other images) metric, that can be computed based on the data extracted from the image. Our practical goal was to find a limited set of digital image features which can be used to create numerical HPIQ models.

Up to this point, the problem of perceived image quality was given very scarce treatment. Most relevant studies were done in the area of digital medicine, using single quality measures such as non-reference image quality metrics for structural MRI [2], a non-reference blur image quality measure based on wavelet transform [3], and information entropy measure for evaluating image quality [4]. There were a few studies attempting multidimensional quality evaluation such as radiation dose and image-quality assessment in computed tomography [5].

To give this problem a full treatment independent on the image content, we chose a much broader and unexplored approach, focusing on multivariate HPIQ modeling. To exclude contextual image bias, we experimented with two completely different sets of digital images: a set of computed tomography scans (MedSet) and a set of scenic forest images (TreeSet). For HPIQ modeling, we chose two different model types, thus making sure that the selection of the most important HPIQ metrics is not influenced by the choice of the model.

Methods and Materials

Image Quality Metrics

To consider all major metrics of digital image quality, we started from an exhaustive list of different non-reference quality-related image metrics described in the previous literature (Table 1). We use $I(x,y)$ to represent image intensity value at the (x,y) pixel; h and w are height and width of image.

Data Preparation

Although our initial interest in HPIQ was driven by medical imaging, it was clear that even within the medical domain, image context can vary significantly. Therefore, to eliminate contextual image bias, we used two image datasets of very different origin: MedSet (CT tomography images, HIPAA-deidentified and IRB compliant under 45 CRF §46.101.(b).(4)), and TreeSet (scenic forest landscapes). Each set contained 50 greyscale images, to exclude color bias. The images were presented in 7000 random pairs to 15 human observers of different age and gender, and without any background in CT imaging. For each image pair, the observers were asked to choose the best image. The task was implemented using Amazon Mechanical Turk (Fig. 1).

Image quality metrics were computed for the original images as well as their lower-resolution versions, to compare Table 1 at different scales. To do so, two lower-resolution copies were produced for each image as first two low-pass levels of the Gaussian pyramid. The low-resolution images were not shown to human observers and were used only for computing the low-resolution metrics. This resulted in a total of 57 image quality measurements per each image (Table 2), to be used as quality-predicting variables.

Modeling HPIQ with Absolute Quality Ranks: Linear Regression Model

Our first approach to HPIQ modeling was based on the assumption that digital image quality can be expressed as a single *quality index* number. Therefore, using the original pairwise image comparison results, we computed a quality index value for every image as the number of this image's wins divided by the number of all comparisons with this image.

Note that in some instances, an image with a higher quality index might have been perceived as inferior when compared with some lower-quality image. An obvious example would be when image A is perceived as better quality than image B, B—as better than C, but C is perceived as better than A. We called image pairs such as (A,C) *inverted*. This non-transitivity in image scores means that one cannot accurately model all human scores with a sequential, single-value outcome. Overall, 10% of pairs were found to be inverted in MedSet, and 14% in TreeSet.

The inverted pairs were not excluded from the visual scoring for two principal reasons. First, we believe that many humans will not be able to perceive image quality in a strictly transitive order. Second, the inverted pairs mean that some images in the dataset are harder to compare, which should provide some valuable information for model building.

Using linear quality indices as a target variable, we then attempted to predict them with a linear regression model. We considered all possible regression models containing various combinations of at most 57 features (Table 2). For each model size $k \leq 57$, we determined the best model with exactly k features, as the model with the least error. To avoid suboptimal stepwise selection, we decided to use an exhaustive search through millions of possible models (feature combinations); branch-and-bound algorithm was applied to speed-up the search process. We observed that model prediction quality essentially stopped to improve after using more than five predictors. Therefore, to avoid overfitting and to maintain model clarity, we limited further analysis to the models with $k \leq 5$ features only.

Figure 2 visualizes our results, plotting model prediction quality (as R squared) for both MedSet and TreeSet datasets combined into a single merged set. Circle color corresponds to

Table 1 Major metrics of digital image quality

Description	Formula
Blur: F_{blur1}, F_{blur2} [6], [7]	
Blur can be detrimental to perceived image quality, and we considered two different blur measures. The first one, F_{blur1}, described in [6], uses a low-pass filter and is based on the principle that gray level of neighboring pixels in a less blurred image changes with higher variation than in its blurred version. The second measure, F_{blur2}, is based on edge extraction using intensity gradient was presented in [7]. The	$D_{ver}(x,y) = \lvert I(x,y) - I(x-1,y)\rvert, \quad x = 2 \dots w, y = 1 \dots h,$ $D_{HOR}(x,y) = \lvert I(x,y) - I(x,y-1)\rvert, \quad x = 1 \dots w, y = 2 \dots h,$ $V_{ver}(x,y) = max\left(0, D_{ver(x,y)} - D_{B\_ver(x,y)}\right), \quad x = 1 \dots w-1, y = 0 \dots h-1,$ $F_{blur\_ver} = \frac{\sum_{x,y=1}^{w-1,h-1} D_{ver(x,y)} - \sum_{x,y=1}^{w-1,h-1} V_{ver(x,y)}}{\sum_{x,y=1}^{w-1,h-1} D_{ver(x,y)}}$ $F_{blur1} = \max(F_{blur\_hor}, F_{blur\_ver})$
resulting measure of blurriness for the whole image is called inversed blurriness and is computed as a ratio of blurred edged pixels count to edge pixels count.	$D_{hor\_mean} = \frac{1}{wh}\sum_{x=1}^{w}\sum_{y=2}^{h-1} \lvert I(x,y-1) - I(x,y+1)\rvert$ $C_{hor}(x,y) = \begin{cases} D_h(x,y), if\ D_h(x,y) > D_{hor\_mean} \\ 0, otherwise \end{cases},$ $E_{hor}(x,y) = \begin{cases} 1, if\ C_{hor(x,y)} > \max\left(C_{hor(x,y-1)}, C_{hor(x,y+1)}\right) \\ 0, otherwise \end{cases},$ $Blur_{hor}(x,y) = \frac{\left\lvert I(x,y) - \frac{1}{2}\left((I(x,y+1) + I(x,y-1)\right)\right\rvert}{\frac{1}{2}\left((I(x,y+1) + I(x,y-1)\right)}$ $B(x,y) = \begin{cases} 1, if\ \max\left(Blur_{hor}(x,y), Blur_{ver}(x,y)\right) > 0.1\ and \\ 0, otherwise \end{cases},$ $F_{blur2} = 1 - \left(\sum_{x,y=1}^{w,h} B(x,y) \Big/ \sum_{x,y=1}^{w,h} E(x,y)\right)$
Image Entropy F_{ent} [8]	
Shannon entropy [8] was computed for the entire image, its foreground, and its background; $p(I_k)$ is the probability of the particular intensity value I_k. We assume that higher entropy means more information is contained within the image, which may affect human perception.	$F_{ent} = -\sum_{k=1}^{n} p(I_k) * \log_2 p(I_k),$
Separability F_{sep} [9]	
Presented in [9], image separability measure F_{sep} shows to what extent various areas of an image can be visually separated. We use the simplest yet most intuitive implementation comparing two major segments: image background (seg1) and foreground (seg2). Average image intensity was computed. All pixels with lower intensity were classified as background, higher – as foreground. To compute separability measure, average difference U for neighboring pixels in a 3x3 sliding window was computed for each image segment, leading to W. Then B is the inversed sum of squared differences of average intensities, and separability F_{sep} is computed as shown	$U_{seg}(x,y) = \sum_{m,n=-1}^{1} (I(x,y) - I(x-n,y-m))^2$ $W = \frac{1}{N_{seg1}}\sum_{(x,y)\in seg_1} U_{seg1}(x,y) + \frac{1}{N_{seg2}}\sum_{(x,y)\in seg_2} U_{seg2}(x,y)$ $B = \left(\frac{1}{N_{seg1}}\sum_{(x,y)\in seg_1} U_{seg1}(x,y) - \frac{1}{N_{seg2}}\sum_{(x,y)\in seg_2} U_{seg2}(x,y)\right)^{-2}$ $F_{sep} = 1000W + B$
Flatness F_{flat} [2]	
Flatness is described in [2] and uses a two-dimensional discrete Fourier transform of the image. The Fourier transform is converted to one-dimensional vector F_V. Next, spectral flatness S_F is computed as ratio of geometric to arithmetic mean, leading to the value of F_{flat}.	$S_F = \left(\prod_{k=1}^{wh} \lvert F_v(k)\rvert^2\right)^{\frac{1}{wh}} \Big/ \frac{1}{wh}\sum_{k=1}^{wh} \lvert F_v(k)\rvert^2$ $F_{flat} = S_F * 1/wh \sum_{x=1}^{w}\sum_{y=1}^{h} \lvert I(x,y) - \overline{I}\rvert^2$
Sharpness F_{sharp} [10]	
This measure [10] is based on the assumption that differences of neighboring pixels change more in the areas with sharp edges. Each pixel is treated as sharp if the sharpness S_{ver} exceeds 0.0001. The number of vertical sharp pixels N_{Sver} is computed, as well as the number of the edge pixels N_{Ever} (found with Canny filter); then the same process is repeated in the horizontal direction.	$\Delta D_2(x,y) = [I\ (x+2,y) - I\ (x,y)] - [I\ (x,y) - I\ (x-2,y)],$ $S_{ver}(x,y) = \sum_{x-t=<k=<x+t} \lvert \Delta D_2(k,y)\rvert \Big/ \sum_{x-t=<k=<x+t} \lvert I(k,y) - I(k-1,y)\rvert$ $F_{sharp} = \sqrt{\left(\frac{N_{S_{hor}}}{N_{E_{hor}}}\right)^2 + \left(\frac{N_{S_{ver}}}{N_{E_{ver}}}\right)^2}$
Blockness F_{block} [11]	
F_{block} [11] estimates image block artifacts, often due to lossy data compression. Absolute intensity differences D for neighboring pixels are obtained for vertical and horizontal directions, then normalized. One-dimensional Discrete Fourier Transform (DFT) $Mhor$ is applied to the horizontal image profile $Phor$, and the same is done for the vertical profile. Due to DFT nature, $Mhor$(T) will have peaks at T=b (w-1)/Z, where b=1,2…Z. Values for $Mhor$(T) at these peak points correspond to horizontal blockness $Blhor$; vertical blockness is computed similarly. In our study, 2, 4, 6 and 8 pixels were used as block width. F_{block}	$D_{hor_{norm}}(x,y) = D_{hor}(x,y) \Big/ \sqrt{\frac{1}{2N}\sum_{k\in[-N,+N],i\neq 0} D_{hor}^2(x+k,y)}$ $P_{hor}(y) = \frac{1}{h}\sum_{x=0}^{w-1} D_{hor_{norm}}(x,y)$ $M_{hor}(T) = \left\lvert \sum_{x=0}^{w-2} P_{hor}(y) e^{\frac{-j2\pi yT}{w-1}} \right\rvert$ where $0 \le T \le w\text{-}2$.

Table 1 (continued)

will be higher for images distorted with block artifacts.	$$Bl_{hor} = 1/M_{hor}(0)\sqrt{1/Z-1}\sum_{b=1}^{Z-1} M_{hor}^2\left(b\frac{w-1}{Z}-1\right)$$ $$F_{block} = \sqrt{rBl_{ver}^2+(1-r)Bl_{hor}^2},\quad r=0.5$$
Fractal Dimension F_{rac} [12]	
Fractal dimensions are computed from the main image contours extracted with a Canny filter, using standard box-count approach. We hypothesize that higher values measure of fractal dimension would correspond to more complex images, thus affecting perceived image quality.	$$\mathrm{F_{frac}} = \lim_{\varepsilon\to 0}\frac{log(N(\varepsilon))}{\log\left(\frac{1}{\varepsilon}\right)}$$
Noise F_{noise} [13]	
Image noise has been the most popular single metric used to assess digital image quality. We included an advanced noise measure developed in [13]. In this work, noise level is described as standard deviation of the Gaussian noise, computed with a patch-based algorithm. All image patches p_i are treated as data in Euclidean space, and their covariance matrix π is defined as shown, where b is number of patches, and μ is the pixel value average of p_i. Then the direction of minimal variance is computed as the π's eigenvector corresponding to the minimal eigenvalue λ_{min}, where ϕ is covariance matrix for noise-free patches. The authors suggest selecting weak textured patches from noisy images because such patches span low-dimensional space. The minimum eigenvalue of their covariance matrix is close to zero, so the noise level F_{noise} can be estimated from as λ_{min} of π', the covariance matrix for weak textured patches.	$$\pi = \frac{1}{b}\sum_{j=1}^{b}(p_i-\mu)(p_i-\mu)^T$$ $$\lambda_{\min}(\pi) = \lambda_{\min}(\phi)+\sigma_n^2$$ $$F_{noise} = \sigma_n^2 = \lambda_{\min}(\pi')$$
Average Gradient F_{AG}**, Edge Intensity** F_{EI} **, Contrast** F_C**, Average Intensity** F_{AI} [14]	
F_{AG} and F_{EI} are taken from [14]. Average intensity F_{AI} and overall image contrast F_C, as well as contrast per pixel F_{CPP} were computed as shown.	$$F_{AG} = \frac{1}{(w-1)*(h-1)}\sum_{x=1}^{w-1}\sum_{y=1}^{h-1}\sqrt{\frac{1}{2}\left(I(x,y)-I(x+1,y)\right)^2+\left(\left(I(x,y)-I(x,y+1)\right)\right)^2}$$ $$G_{hor}(x,y) = \lvert I(x+1,y-1)+2I(x+1,y)+I(x+1,y+1)\rvert - \lvert I(x-1,y-1)+2I(x-1,y)+I(x-1,y+1)\rvert$$ $$F_{EI} = \sum_{x}^{w}\sum_{y}^{h}\left({G_{hor}}^2(x,y)+{G_{ver}}^2(x,y)\right)$$ $$F_{AI}(I) = \frac{1}{hw}\sum_{x,y=1}^{w,h} I(x,y)$$ $$F_C = \frac{I_{max}-I_{min}}{I_{max}+I_{min}}$$ $$F_{CPP}(I) = \frac{F_C}{hw}$$

specific model size k, and circle size—to the model error. Thus, the largest circles in the left bottom corner correspond to models of size $k=1$. You can see that with increased k, model error decreases, and models converge to the right top corner.

One can also observe that the circles on the Fig. 2 plot tend to cluster along the diagonal line, which means that most HPIQ models perform similarly on both MedSet and TreeSet. Moreover, the more model features k are used, the closer circles approach to the right top corner. As a result, one can arrive to another interesting conclusion: higher model size k corresponds not only to the more accurate, but also to the more context-independent models, capturing the true nature of HPIQ regardless of the image content.

Figure 3 illustrates similar results obtained for MedSet and TreeSet independently. As the figure indicates, the models selected as the best for one dataset performed well on the other. Despite the obvious differences between the CT scans and forest landscapes, models optimal for one set were among the best performers for the other, which again confirms our hypothesis of context-independent HPIQ modeling.

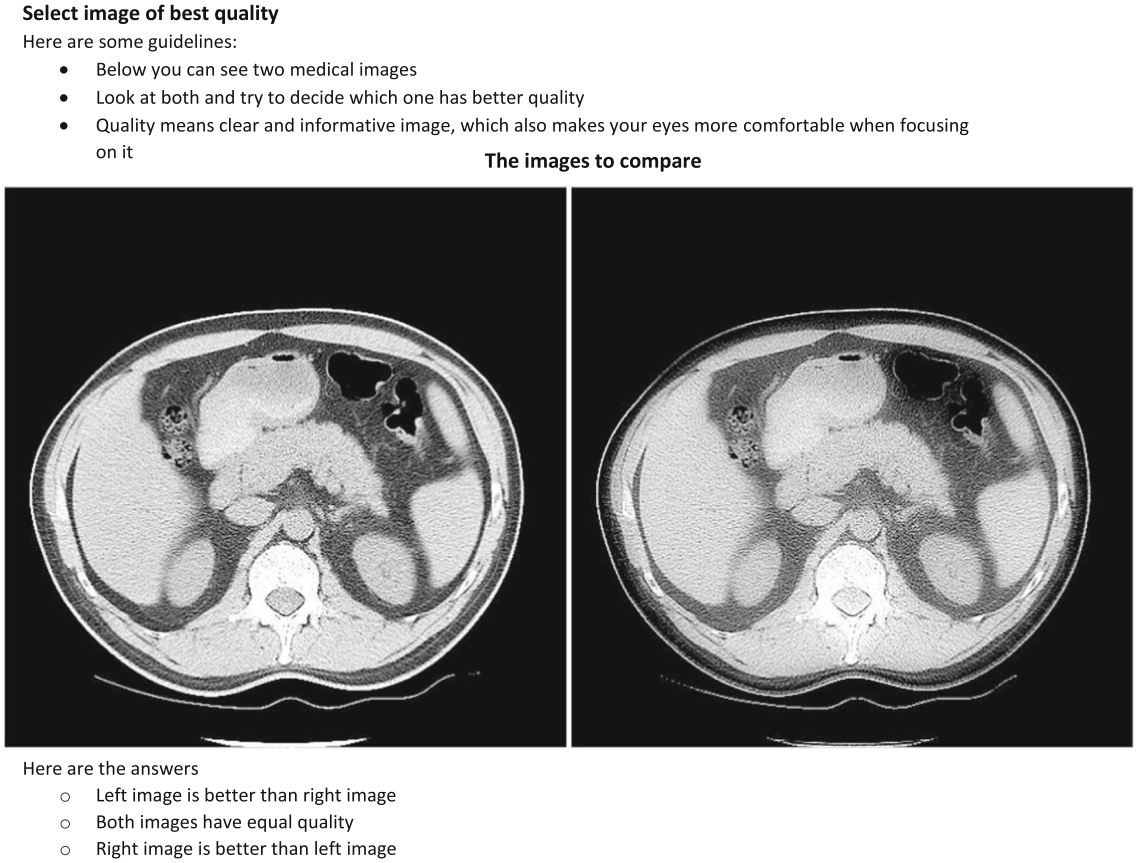

Fig. 1 Amazon Mechanical Turk assignment for image markup

Table 3 summarizes the best predictors selected for each number of features defined in Tables 1 and 2. It provides us with some significant insights. As one can see, HPIQ can be captured with a very limited set of the most principal digital image metrics. It can be assumed that these metrics play the key role in our perception of the image quality. One can see that image entropy, blur, and blockness at different resolution levels turned out to be the most prominent predictors of HPIQ.

The number of inverted pairs computed for predicted quality measures in comparison to initial matrix of comparisons

Table 2 Correspondence between described measures and names of variables used in HPIQ models. Indices 0, 1, and 2 correspond to three levels of Gaussian pyramid; the first index in blockness corresponds to the block size

Measure	Name	Corresponding variables
Blurriness 1	F_{blur1}	blur10, blur11, blur12
Blurriness 2	F_{blur2}	blur20, blur21, blur22
Shannon entropy	F_{ent}	ent10, ent11, ent12; entb0 (background), entf0 (foreground)
Separability	F_{sep}	sep0, sep1, sep2
Flatness	F_{flat}	flat0, flat1, flat2
Sharpness	F_{sharp}	sharp0, sharp1, sharp2
Blockness	F_{block}	block20, block40, block60, block80, block21 etc.
Fractal dimension	F_{frac}	frac0, frac1, frac2
Noise	F_{noise}	noise0, noise1, noise2
Average gradient	F_{AG}	ag0, ag1, ag2
Edge intensity	F_{EI}	ei0, ei1, ei2
Contrast	F_{C}	contr20, contr21, contr22
CPP—contrast per pixel	F_{CPP}	contr10, contr11, contr12
Average intensity	F_{AI}	intens0, intens1, intens2

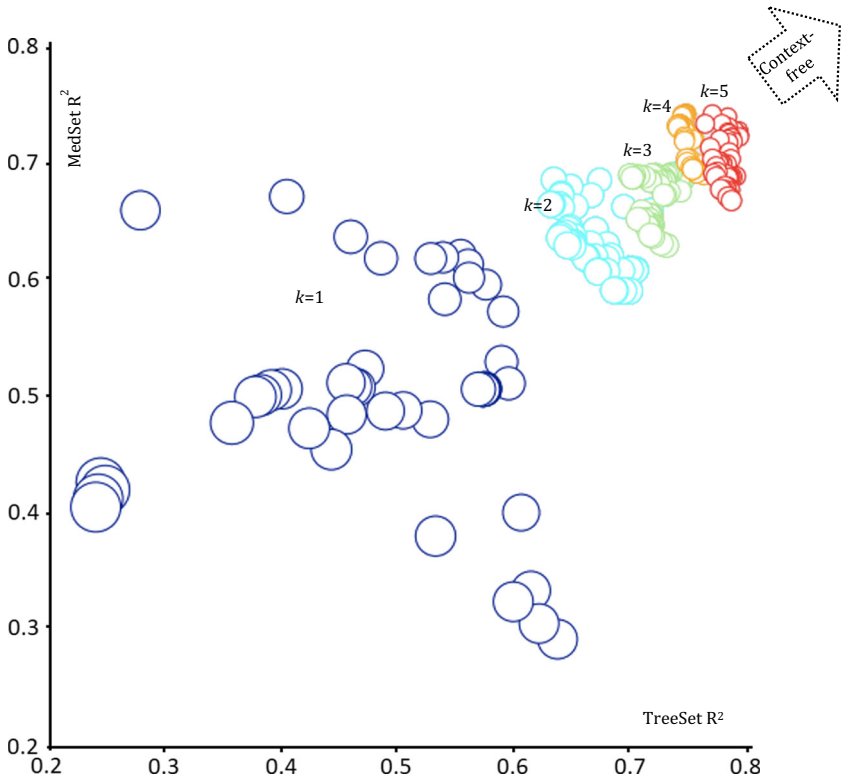

Fig. 2 Optimizing quality-predicting models for both MedSet and TreeSet data. Navy, blue, green, orange, and red circles correspond to modal sizes *k* from 1 to 5, respectively. Note that as model accuracy increases (with model size *k*), the models approach the right top corner, thus becoming equally accurate for the two sets of images

was 30% for MedSet and 29% for TreeSet, thus resulting in a 70–71% prediction. Although it was not matching, the original counts of the inverted pairs (10 and 14% respectively), given the small model size and HPIQ complexity, we consider this as a very good result.

Modeling HPIQ with Relative Quality Ranks: Nonlinear Elo Model

In our previous approach, the original observations were reduced to sequential quality indices, to be used as a target for a

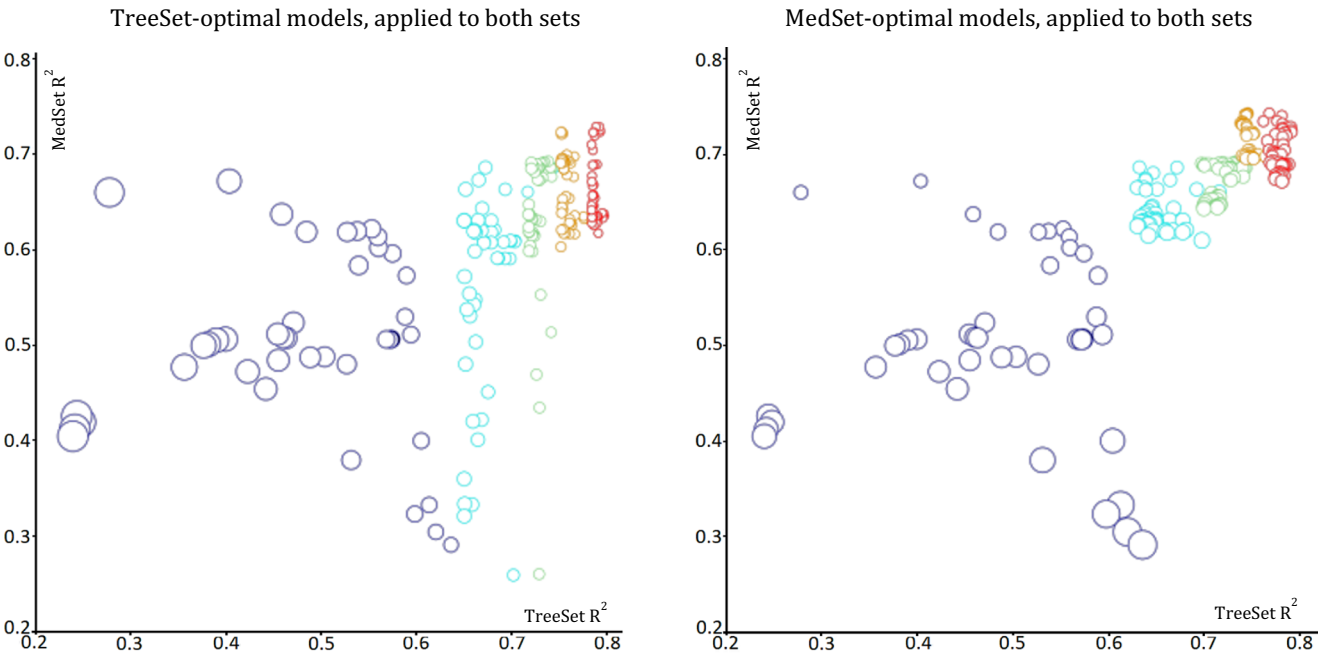

Fig. 3 Optimizing quality-predicting models for TreeSet (left) and MedSet (right)

Table 3 Best predictor values for models with restricted sets of metrics. Table contains best three models according to average error on two datasets

Model size k	Best predictors for both datasets	Best predictors for TreeSet	Best predictors for MedSet
1	- blur10 - blur12 - sep0	- ag1 - sharp1 - ei1 or ei0	- ent10 - ent11 - sep0
2	- blur20, sep0 - blur20, sep1 - entf1, blur20 - blur20/21, intens0/1/2	- blur20, entf0 - entb0, block60 - entb0, frac0	- blur20, sep0 - blur20, sep1 - blur20, intens0
3	- blur20, entb0, sharp1 - blur10, blur11, blur22 - blockness measures + blur - blur20, entb0, frac0	- blur20, entb0, frac2 - entb0, sep0, flat2 - blur20, block22, block62	- blur10, blur11, blur22 - contr20, blur21, noise2 - contr20, intens0, ent11
4	- blur20 + blockness measures - blur11, entb1, intens1, block22	- entb0, sep0, block80, flat2 - blur10, entb0, sep0, flat2	- block62, blur20, contr10, block22 - blur20, contr20, block62, block22
5	- entb0, blur21, flat1, ei1, frac2 - entb0, blur21, flat1, ei1, block62	- blur10, entb0, sep0, block40, flat2 - blur10, entb0, sep0, block80, flat2	- blur20, block60, block62, block22 - blur20, ei0, ei1, block22, block42

linear regression model. This reduction eliminated most of the original comparison pairs. To overcome this limitation and to use our experimental data to its full extent, we studied another quality-rating approach, where all original pairwise comparisons can be used to discover the best predictive features.

The approach is based on the Elo rating system for chess tournaments [15]. Each image pair presented to an observer is considered as an independent Bernoulli test where each outcome (such as winning of image A over image B) has its own probability. Image comparison outcomes are still determined by individual image ratings, so that the image with higher rating wins. Rating of image A is modeled a linear combination of its k features F_j with weights w_j:

$$R_A = \sum_{j=1}^{j=k} w_j F_j$$

The probability of image A rating being higher than image B rating is modeled as a logistic function:

$$P_i(R_A > R_B) = \frac{\exp(R_A - R_B)}{1 + \exp(R_A - R_B)}$$

The optimal set of features weights w_j would correspond to the optimal Elo model explaining the observed comparisons. Outcome x of each comparison can be 0 or 1, which can be written using Bernoulli formula as follows:

$$P_i(R_A > R_B) = P_i(x) = P_i^x (1 - P_i)^{1-x}, x = \{0, 1\}$$

Finally, likelihood function is written as the following product:

$$L = \prod_{i=1}^{N} \left[P_i^x (1 - P_i)^{1-x} \right]$$

To obtain image rankings that would produce pairwise comparisons closest to the original (human observer) comparison data, one should iteratively train features weights w_j to maximize the logarithm of likelihood L. This method was applied to various combinations of five features discovered in our previous method, for TreeSet and MedSet datasets, independently. Then best models for combined set of all images were also obtained. In case of testing model on both MedSet and TreeSet datasets, we used the sum of log

Table 4 Best predictor values for models with restricted sets of metrics; model size $k = 5$. Table contains best six models according to average rate of correct pairwise comparisons

Model with size $k = 5$	Ratio of correct pairwise comparisons predicted by the model		
	Both sets combined	MedSet only	TreeSet only
block21, block22, blur20, ent10, ent11	0.69	0.75	0.74
block82, ent11, ent12, entb0, entf0	0.69	0.73	0.75
contr12, contr20, ent11, entf2, sharp2	0.69	0.76	0.73
block42, block62, ent11, intens2, sharp2	0.67	0.63	0.67
block20, block22, ent10, entf1, noise1	0.66	0.74	0.75
block20, block22, ent10, entf1, noise1	0.66	0.75	0.76

likelihood for two datasets separately, and took the average of features weights for the two datasets. To compare models, we used the rate of truly detected pairwise outcomes as are presented in Table 4 (for model size $k = 5$).

Looking at the Table 4 models, one can observe that the most important HPIQ predictors identified by the Elo model approach correspond to the ones found with the linear regression HPIQ model in Table 3. Thus, the entropy of the whole image, and the entropy of the image background and foreground on all levels of Gaussian pyramid are still the most essential predictors of HPIQ. In addition, blurriness, blockness, noise, sharpness, and contrast are also present in the top winning models. However, due to a less-constrained modeling, Elo approach achieves 24–27% of inverted pairs on separate sets, which is better than with linear regression.

Conclusions

Using two image datasets of different origin, we identified the most important metrics responsible for the human perception of image quality (HPIQ). The analysis was performed with two unrelated HPIQ-modeling algorithms, to eliminate dependency on the model.

The first algorithm used an absolute quality index that was obtained from the initial comparisons as a ratio of pairwise wins; linear regression was used as the predictive model. The second algorithm modeled a raw pairwise HPIQ comparisons matrix with nonlinear Elo model. Comparing the two algorithms based on their fraction of falsely predicted pairwise comparisons (inverted image pairs), we obtained 29–30% for the first, and 24–27% for the second approach. This result indicates that at least 70% of all pairwise comparisons can be predicted with a rather simple numerical model and only up to five key image metrics. Given that the original data already contained 10–14% of inverted (unpredictable) pairs, we find our result extremely interesting and satisfactory.

Both HPIQ models led to the unification of the most influential predictors (metrics), and both models resulted in similar predictor sets. It is particularly visible in Figs. 2 and 3, where, as model accuracy increases, the best MedSet and TreeSet models converge to each other. We view this as the most interesting result of our study, demonstrating that human perception of image quality is largely context-independent, and therefore can be efficiently quantified with an objective, numerical model. As a result, one can build concise HPIQ models to automatically evaluate digital image quality as it is perceived by the humans.

References

1. Geyer LL, Schoepf J, Meinel FG, Nance JW, Bastarrika G, Leipsic JA, Paul N, Rengo M, Laghi A, De Cecco CN: State of the Art: Iterative CT Reconstruction Techniques. Radiology 276(2):339–357, 2015
2. Jeffrey MPC-S, Woodard P: No-Reference image quality metrics for structural MRI. Neuroinformatics 4, 2006
3. Serir A, Kerouh F, A no-reference blur image quality measure based on wavelet transform, Digital Information Processing and Communications, 2012.
4. Lee Y, Matsuyama E, Tsai DY: Information Entropy Measure for Evaluation of Image Quality. Journal of Digital Imaging 21:338–347, 2008
5. ICRU: Radiation Dose and Image-Quality Assessment in Computed Tomography, vol. 12, O. U. Press, Ed., Journal of the ICRU, 2013.
6. Crété-Roffet F, Dolmiere T, Ladret P, Nicolas M: The Blur Effect: Perception and Estimation with a New No-Reference Perceptual Blur Metric. Grenoble: SPIE Electronic Imaging Symposium Conf Human Vision and Electronic Imaging, États-Unis d'Amérique, 2007
7. Choi MG, Jung JH, Jeon JW: No-Reference Image Quality Assessment using Blur and Noise. *International Scholarly and Scientific Research & Innovation* 3(2):184–188, 2009
8. Information theory, Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Entropy_(information_theory). [Accessed 01 02 2018].
9. De K: A new no-reference image quality measure to determine the quality of a given image using object separability, Taipei: Machine Vision and Image Processing (MVIP), 2012 International Conference on, 2012.
10. Chen F, Doermann D, Kumar J: Sharpness estimation for Document and Scene Images, in *Pattern Recognition (ICPR), 2012 21st International Conference on*, Tsukuba, 2012.
11. Chen JBC: A blind reference-free blockiness measure, Shanghai: in Proceedings of the Pacic Rim Conference on Advances in Multimedia Information Processing: part I, 2010
12. Fractal Dimension, Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Fractal_dimension. [Accessed 01 02 2018].
13. Tanaka M, Okutomi M, Liu X: Noise Level Estimation Using Weak Textured Patches of a Single Noisy Image, IEEE International Conference on Image Processing (ICIP), 2012.
14. Zheng X, Hu X, Zhou W, Wang W, Yuan T: A method for the evaluation of image quality according to the recognition effectiveness of objects in the optical remote sensing image using machine learning algorithm. PLoS ONE, 2014
15. Elo AE: 8.4 Logistic Probability as a Rating Basis. The Rating of Chessplayers, Past & Present. NY: Press International, 2008